SEQUENCES OF PROTEINS OF IMMUNOLOGICAL INTEREST

FIFTH EDITION

Tabulation and Analysis of
Amino Acid and Nucleic Acid Sequences of Precursors,
V-Regions, C-Regions, J-Chain, T-Cell Receptors for Antigen,
T-Cell Surface Antigens, β_2-Microglobulins,
Major Histocompatibility Antigens, Thy-1, Complement,
C-Reactive Protein, Thymopoietin, Integrins, Post-gamma Globulin,
α_2-Macroglobulins, and Other Related Proteins

1991

Elvin A. Kabat, Tai Te Wu*, Harold M. Perry†,*
Kay S. Gottesman‡, and Carl Foeller‡

*Depts. of Microbiology, Genetics and Development, and Neurology, Cancer Center/Institute of Cancer Research, College of Physicians and Surgeons, Columbia University, New York, NY 10032 and the National Institute of Allergy and Infectious Diseases, and the Office of the Director National Institutes of Health, Bethesda, MD 20892.

*Depts. of Biochemistry, Molecular Biology, and Cell Biology, and Engineering Sciences and Applied Mathematics and Biomedical Engineering, Northwestern University, Evanston, IL 60208 and the Cancer Center, Northwestern University Medical School, Chicago, IL 60611

†BBN Systems and Technologies, 10 Moulton Street, Cambridge, MA 02138

‡Formerly with BBN. Present address Laboratory for Applied Research in Academic Information, William H. Welch Medical Library, The Johns Hopkins University, Baltimore, Md 21205

The collection and maintenance of this data base is sponsored under grant 5R01 AI-125616 to E.A. Kabat of Columbia University by the following components of the National Institutes of Health, Bethesda, MD 20892:

Office of the Director
National Center for Research Resources
National Cancer Institute
National Institute of Allergy and Infectious Diseases
National Institute of Diabetes, Digestive and Kidney Diseases
National Institute of General Medical Sciences
National Library of Medicine

Work with the PROPHET software package is supported by a subcontract from Columbia University to BBN Systems and Technologies, Cambridge, MA 02138

U.S. DEPARTMENT OF HEALTH
AND HUMAN SERVICES

Public Health Service
National Institutes of Health

NIH Publication No. 91-3242

end of each table and the summary tables, to evaluate the probability that a given amino acid at a given position may not be correct. This is most readily done for the framework residues of the V-region and for the C-region; in the complementarity-determining regions this is more difficult because of the high variability.

AMINO ACID SEQUENCES

The first column in each table gives the residue number. Except for complement, T-cell surface antigens, integrins and miscellaneous proteins, the second column is a tabulation of invariant residues. Since exceptions to invariance are found, the frequency, if less than 1.0 and greater than or equal to 0.95, is indicated alongside the residue listed as invariant; when only a single sequence is available, this is not given. These rows are shaded in grey. Each sequence is tabulated in each subsequent column. Three dashes (---) indicate that no amino acid is present at that position and that the sequence continues. In all instances residues considered uncertain by the authors have not been included in the table. In some instances the symbol # is used to indicate that several amino acid residues were found in one position, and these residues are listed in the notes. The four columns at the end of each table give:

1. the number of residues sequenced at that position,
2. the number of different amino acids found at that position,
3. the number of times the most common amino acid occurred and that amino acid in parentheses, and
4. the variability.

These columns are included only in tables with more than five sequences. Miscellaneous tables have only columns corresponding to the first two above.

Variability is calculated (16) as:

$$\text{Variability} = \frac{\text{Number of different amino acids occurring at a given position}}{\text{Frequency of the most common amino acid at that position}}$$

An invariant position would have a variability of one; if 20 amino acids occurred with equal frequency, the variability would be 20 divided by 0.05 equals 400. If, for example, four different amino acids Ser, Asp, Pro, and Thr occurred at a given position, and of 100 sequences available at that position, Ser occurred 80 times, the variability would be 4/0.8 = 5. When any of the amino acid residues, sequenced directly as amino acids, were not identified completely and are listed as Glx (or Asx), two values, separated by a comma, are given in the last three columns. The first value in each of these columns is calculated assuming that only one of the two possibilities, e.g., Glu or Gln (or Asp or Asn) occurred, while the second considers that both were present and maximizes variability. In the variability plots, the horizontal bars indicate the two values.

When two or more amino acids are most common and occur with equal frequency, they are tabulated as a note, and the symbol + is used in the next to last column. If no sequence data have been reported for any position, there are no entries in the last four columns. Variability is not calculated for insertions or if only a single sequence is known. When the translated sequence of a clone corresponds to a previously listed sequence of a plasmacytoma from

which it was prepared, only one sequence is listed so that the variability computations are not affected, and a note is included. If a given sequence is associated with any antibody activity, this is indicated by an asterisk alongside the protein heading, and the antibody specificities are given in a separate list with binding constants if available. The notes list the a-allotypes for the rabbit heavy chain V-region and the b-allotypes for the constant domain of the rabbit kappa light chain. A key reference to the sequence is given; generally the most recent reference since it is usually the most nearly complete, but often several references are included, especially when revisions of a sequence have been made. Notes are of two types: general notes about a table indicated by the symbol #, and specific notes indicated by the sequence number.

Signal Sequences

The signal (precursor) amino acid sequences of immunoglobulin chains are listed as human, mouse, and miscellaneous for kappa light chains, for lambda light chains, and for heavy chains for a total of nine precursor tables. They were obtained either by direct sequencing of signal proteins (17-19) or by translating nucleotide sequences from DNA clones. Signal segments range from 17-29 amino acid residues in length and are thus numbered from -29 to -1. Genomic DNA clones contain introns of varying length that interrupt the coding sequence of the precursor within the codon for position -4, and in rare cases for position -6. Thus, the L-gene encodes the leader peptide to position -4 and the 5' end of the V-gene codes for positions -4 to -1.

The signal amino acid sequences of the T-cell receptors for antigens, β2-microglobulins, major histocompatibility complex proteins, complement components, integrins, and other related proteins are listed in separate tables.

By conformational energy calculations, the core V_κ hydrophobic Leu-Leu-Leu-Trp-Val-Leu-Leu-Leu (MOPC321, MOPC63) exists in an alpha helical conformation, terminated by chain reversal conformations in the four C-terminal residues Trp-Val-Pro-Gly; the four amino terminal residues are compatible with the alpha helix (20).

Variable Region Sequences

The variable regions (21) of immunoglobulins have been shown to contain hypervariable segments in their light (16,22-26) and heavy (27-30) chains, of which certain residues have been affinity labeled with haptenic determinants (31-44). Three hypervariable segments of light chain were delineated from a statistical examination of sequences of human V_κ, human V_λ, and mouse V_κ light chains aligned for maximum sequence similarity (16,23,24,27). These and the three corresponding segments of the heavy chains (27) were hypothesized (16,27) to be the complementarity-determining regions or segments (CDR) containing the residues which make contact with various antigenic determinants, several years before high resolution x-ray structures were determined, and this has now been verified by X-ray diffraction studies at high resolution for all antibodies examined Figures 3-47. The proposed fourth hypervariable region (cf. 30) of heavy chains is not part of the antibody combining site (27). The rest of the V-region constitutes the framework (16,27,45-54). It is convenient to identify the framework segments (FR1, FR2, FR3, and FR4) and the complementarity-determining segments (CDR1, CDR2, and CDR3) with the three CDRs separating the four FRs. The CDRs in the stereo Figures 3-47 have solid circles for each residue. References and comments are given with each figure and are not listed in the bibliography. The residue numbers for these segments are given in Table I.

TABLE I

Amino Acid Residues Associated with Framework(FR) and Complementarity Determining Regions (CDR) of the Variable Domains of Immunoglobulin Light (V_L) and Heavy (V_H) Chains

Segment	Light Chain	Heavy Chain
FR1	1–23 (with an occasional residue at 0, and a deletion at 10 in V_λ chains)	1–30 (with an occasional residue at 0)
CDR1	24–34 (with possible insertions numbered as 27A,B,C,D,E,F)	31–35(with possible insertions numbered as 35A,B)
FR2[a]	35–49[a]	36–49
CDR2	50–56	50–65 (with possible insertions numbered as 52A,B,C)[b]
FR3	57–88	66–94 (with possible insertions numbered as 82A,B,C)
CDR3	89–97 (with possible insertions numbered as 95A,B,C,D,E,F)	95–102 (with possible insertions numbered as 100A,B,C,D,E,F,G,H,I,J,K)
FR4	98–107 (with a possible insertion numbered as 106A)	103–113

[a] Five Basilea rabbits (λ) immunized with type II pneumococci and which produced anti-type II pneumococcal polysaccharide had Met at position 48 and an insertion of four amino acid residues between positions 48 and 49; in four of the five the sequence was Glu, Leu, Lys, Ser and the fifth was Trp, Leu, Arg, Lys (53,54,63,64); the others were not sequenced at these positions (for references see table of rabbit λ amino acid sequences.)

[b] In the rabbit, Mage et al. (65) consider position 65 in V_H to be in FR3, since it is allotype related.

The V-genes for the light chains code to amino acid position 95, and the J-minigenes from position 97 to 107 for lambda and 108 for kappa light chains. Position 96 is usually the site of V-J joining by recombination and may be coded partly by the V-gene and partly by the J-minigene. Because the site of V-J recombination could occur at different positions within a codon, different amino acid residues may result at this position. We have changed the location of the inserted residues from 97A-F (2) to 95A-F, since it makes for better alignment by confining chains of different lengths to the V-gene region. In mouse V_κ chains, J1 and J2 were used 5 to 10 times more frequently than J4 and J5 (55).

The V-genes for the heavy chains code up to amino acid position 94 and are followed by the D- and J-minigenes. Because of the extensive variation in the lengths of D-minigenes, and their ability to be read in different reading frames (56), the exact boundary between D and J is not always located at the same amino acid position. In addition, the lengths of the J encoded amino acid sequences vary by a few amino acid residues. Moreover, the process of D-J joining appears to involve insertions of extra nucleotides between V and D and between D and J, termed the N region (57-61) and correlates with the appearance of terminal deoxytransferase in B cells (60). The original numbering system for the heavy chains has therefore been retained. Wysocki et al. (61) have provided some evidence suggesting a non-random origin for the V_H-D_H junction, perhaps a minigene, rather than random addition of the N nucleotides. Light chains do not appear to have N sequences at the V_L-J_L junction (62), but show an additional residue 95A which probably results from V_L-J_L joining. N sequences are generally rare in fetal and neonatal mouse V_H-D-J_H junctions (62), only 1/87 DNA and 17/146 RNA sequences contained N regions, an incidence much lower than in adults indicating that N insertion is developmentally regulated both in T and B cells. P elements also contribute to diversity but are templated (62a).

In the tables of V-regions, the FR and CDR are separated by horizontal lines for convenience in reading. One mouse kappa light chain, MPC 11, has an extra segment of 12 amino acid residues between position 1 and the signal sequence (66). Several chains have internal deletions.